Claims

1. A method of processing digitized textual information, the information being organized in terms, documents and document corpora, where each document contains at least one term and each document corpus contains at least one document, the method comprising:

generating a concept vector for each document in a document corpus, the concept vector conceptually classifying the contents of the document on a relatively compact format, and

generating, for each term in the document corpus, a term-to-concept vector describing a relationship between the term and each of the concept vectors, **characterized by** the term-to-concept vectors being generated on basis of the concept vectors, and the method comprising:

receiving the term-to-concept vectors for the document corpus and on basis thereof generating a term-term matrix describing a term-to-term relationship between the terms in the document corpus, and

processing the term-term matrix into processed textual information.

2. A method according to claim 1, **characterized by** each document in the document corpus being associated with a document-concept matrix representing at least one concept element whose relevance with respect to the document is described by a weight factor, the generation of each term-to-concept vector comprising:

identifying a term-relevant set of documents in the document corpus, each document in the term-relevant set containing at least one occurrence of the term,

calculating a term weight for the term in each of the documents in the term-relevant set,

retrieving a respective concept vector being associated with each document in the term-relevant set where the term weight exceeds a first threshold value,

selecting a relevant set of concept vectors including any concept vector in which at least one concept component exceeds a second threshold value,

calculating a non-normalized term-to-concept vector as the sum of all concept vectors in the relevant set, and

normalizing the non-normalized term-to-concept vector.

3. A method according to any one of the preceding claims, **characterized by** the generation of the term-term matrix comprising:

retrieving, for each term in each combination of two unique terms in the document corpus, a respective term-to-concept vector,

generating a relation vector describing the relationship between the terms in each combination of two unique terms, each component in the relation vector being equal to a lowest component value of corresponding component values in the term-to-concept vectors,

generating a relationship value for each combination of two unique terms as the sum of all component values in the corresponding relation vector, and

generating a matrix containing the relationship values of all combinations of two unique terms in the document corpus.

4. A method according to any one of the preceding claims, **characterized by**

calculating a statistical co-occurrence value between each combination of two unique terms in the document corpus, the statistical co-occurrence value describing a dependent probability that a certain second term exists in a document provided that a certain first term exists in the document, and

incorporating the statistical co-occurrence values into the term-term matrix to represent lexical relationships between the terms in the document corpus.

5. A method according to any one of the preceding claims,

characterized by displaying the processed textual information on a format being adapted for human comprehension.

6. A method according to claim 5, **characterized by** the displaying step involving presentation of at least one of:

at least one document identifier specifying a document being relevant with respect at least one term in a query,

at least one term being related to a term in a query, and

a conceptual distribution representing a conceptual relationship between two or more terms in the document corpus, the conceptual distribution being based on shared concepts which are common to said terms.

7. A method according to claim 6, **characterized by** the displaying step involving presentation of at least one document identifier specifying a document being relevant with respect to at least one term in a query in combination with at least one user specified concept.

8. A method according to claim 7, **characterized by** selecting the at least one user specified concept from the shared concepts in the conceptual distribution.

9. A method according to any one of the claims 5 - 8, **characterized by** illustrating the conceptual relationship between a first term and at least one second term by means of a respective relevance measure being associated with the at least one second term in respect of the first term.

10. A method according to claim 9, **characterized by** displaying the processed textual information on a graphical format which visualizes the strength in the conceptual relationship between at least two terms.

11. A method according to any one of the claims 9 or 10, **characterized by** displaying the processed textual information

as a distance graph in which

each term constitutes a node,

a node representing a first term is connected to one or more other nodes representing secondary terms to which the first term has a conceptual relationship of at least a specific strength, and

the relevance measure between the first term and the at least one second term is represented by a minimum number of node hops between the first term and the at least one second term.

12. A method according to any one of the claims 9 or 10, **characterized by** displaying the processed textual information as a distance graph in which

each term constitutes a node,

a node representing a first term is connected to one or more other nodes representing secondary terms to which the first term has a conceptual relationship, each connection is associated with an edge weight representing the strength of a conceptual relationship between the first term and a particular secondary term, and

the relevance measure between the first term and a particular secondary term is represented by an accumulation of the edge weights being associated with the connections constituting a minimum number node hops between the first term and the particular secondary term.

13. A method according to any one of the preceding claims, **characterized by** each term representing one of:

a single word,

a proper name,

a phrase, and

a compound of single words.

14. A method according to any one of the preceding claims, **characterized by** updating the document corpus with added

data in form of at least one new document by means of:

identifying any added terms in the new document which lack a representation in the document corpus,

identifying any existing terms in the new document which were represented in the document corpus before adding the at least one new document,

retrieving, for each of the existing terms, a corresponding concept vector,

generating a new concept vector with respect to the at least one new document as a sum of the corresponding concept vectors,

normalizing the new concept vector into a normalized new concept vector, and

assigning the normalized new concept vector to each of the added terms in the new document.

15. A computer program directly loadable into the internal memory of a digital computer, comprising software for performing the steps of any of the claims 1 – 14 when said program is run on a computer.

16. A computer readable medium, having a program recorded thereon, where the program is to make a computer perform the steps of any of the claims 1 – 14.

17. A search engine (115) for processing an amount of digitized textual information and extracting data there from, the information being organized in terms, documents and document corpora, where each document contains at least one term and each document corpus contains at least one document, comprising:

an interface (116) adapted to receive a query (Q) from a user, and

a processing unit (150) adapted to process a document corpus on basis of the query (Q) and return processed textual information (R) being relevant to the query (Q), said process

involving

generating a concept vector for each document in the document corpus, the concept vector conceptually classifying the contents of the document on a relatively compact format, and

generating, for each term in the document corpus, a term-to-concept vector describing a relationship between the term and each of the concept vectors,

characterized in that the processing unit (150) in turn comprises:

a processing module (151) adapted to receive the term-to-concept vectors for the document corpus and on basis thereof generate a term-term matrix describing a term-to-term relationship between the terms in the document corpus, and

an exploring module (152) adapted to receive the query (Q) and the term-term matrix, and on basis of the query (Q), process the term-term matrix into the processed textual information (R).

18. A database (130) holding an amount of digitized textual information being organized in terms, documents and document corpora, where each document contains at least one term and each document corpus contains at least one document,

each document in a document corpus being associated with concept vector which conceptually classifies the contents of the document on a relatively compact format, and

each term in the document corpus being associated with a term-to-concept vector describing a relationship between the term and each of the concept vectors,

characterized in that it is adapted to deliver the term-to-concept vectors to a search engine (115) according to the claim 17.

19. A database (130) according to claim 18, **characterized in that** it comprises an iterative term-to-concept engine adapted to receive fresh digitized textual information added to the database (130) and on basis of this information:

generate concept vectors for any added document, and

generate a term-to-concept vector describing a relationship between any added term and each of the concept vectors.

20. A server (110) for providing data processing services in respect of digitized textual information, **characterized in that** it comprises

a search engine (115) according to claim 17, and

a communication interface (112) towards a database (130) according to any one of the claims 18 or 19.

21. A system for providing data processing services in respect of digitized textual information, **characterized in that** it comprises

a server (110) according to claim 20,

at least one user client (120) adapted to communicate with the server (110), and

a communication link (141; 142) connecting the at least one user client (120) with the server (110).

22. A system according to claim 21, **characterized in that** an internet (140) accomplishes at least a part of the communication link (141; 142), and the at least one user client (120) comprises a web browser (121) which in turn provides:

a user input interface (121a) adapted to receive queries (Q) from a user and forward the queries (Q) to the server (110) via the communication link (141), and

a user output interface (121b) adapted to receive processed textual information (R) from the server (110) via the communication link (142) and present the processed textual information (R) to the user.

23. A method of processing digitized textual information, the information being organized in terms, documents and document corpora, where each document contains at least one term and each document corpus contains at least one document, the

method comprising:

identifying a particular document corpus (910),

filtering the identified document corpus wherein a number of documents fulfilling at least one specified criterion are selected (920; 950, 960), and

producing a new document corpus exclusively containing the selected documents (930; 970).

24. A method according to claim 23, **characterized by** the filtering involving:

identifying a number of document clusters in the identified document corpus by means of a document clustering algorithm (920a),

generating, for each identified document cluster, a representative document vector by means of the document clustering algorithm (920b, 920c, 920e), and

removing all non-clustered documents from the identified document corpus (920d).

25. A method according to claim 23, **characterized by** the filtering involving:

receiving a user input specifying at least one of one or more concepts and one or more terms (950),

selecting, from the identified document corpus, documents being related to at least one of the concepts or the terms (960), and

removing all non-selected documents from the identified document corpus.

26. A method according to any one of the claims 23 - 25, **characterized by** the identified document corpus having been processed according to the method according to any one of the claims 1 –14.